JMCS

Journal of Mathematics & Computing Science

Vol 1. Issue 1, May 2016

Universiti Teknologi MARA

Unit Penerbitan UiTM Kelantan

e-ISSN: 0128-2697

The Effect of High Leverage Points on VIF Measures on Noncollinear Data

Nurul Bariyah Ibrahim[1*], Prof. Dr. Habshah Midi[2], Noor Ilanie Nordin[3], Nor Azima Ismail[4], Nur Elini Jauhari[5], Norafefah Mohamad Sobri[6]

[1,3,4,5,6]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kelantan, Bukit Ilmu, Machang, Kelantan, Malaysia
bariyah@kelantan.uitm.edu.my
[2]Faculty Of Science, University Putra Malaysia

Abstract:Multicollinearity is a case of multiple regression in which the predictor variables are highly correlated among themselves. The problem will get more complicated when multicollinearity exists together with high leverage points. The usage of classical VIF for multicollinearity diagnostics is not reliable as it is not resistant to the presence of high leverage points. In this study, we proposed RVIF which is based on the MM estimator in the detection of multicollinearity due to the high leverage point. The computation of RVIF is based on robust coefficient determination which is called RR^2 (MM). We denote this estimator as RVIF (MM). The numerical results and Monte Carlo simulation study indicate that the CVIF performs poorly in the presence of high leverage point and the proposed RVIF is very resistant to the high leverage point and unable to detect the multicollinearity in the data.

Keywords: High leverage point, Multicollinearity, MM Estimator, Robust coefficient determinations, Variance Inflation Factors.

1 Introduction

In multiple regression analysis, the nature and significance of the relations between the predictor variables and response variable are often of particular interest. In the application of regression analysis, multicollinearity is a problem that always occurs when two or more predictors are correlated with each other. Multicollinearity among regressors is an exciting and common property of data. The consequences of multicollinearity for estimation and inference are well known such as unreliable estimation results, high standard errors, coefficients with wrong signs and implausible magnitudes (Belsley et al.[1]). It is well known (Chatterjee and Hadi[2]; Belsley et.al.[1]) that multicollinearity does affect least squares estimates but not the least squares residuals and predictions, in which the tests are based. However, our investigation wants to examine the cause of the multicollinearity problem. Multicollinearity can also affect the parameter estimates and it is used in interpreting a fitted regression model.

The purpose of this study is to identify whether or not high leverage points affect the multicollinearity problem. Specifically, the objectives of this study are:

1. To detect the multicollinearity problem using classical and robust Variance Inflation Factors (VIF).
2. To examine the effect of high leverage point on the classical and robust VIF measure in non collinear data.
3. To compare the performance of the classical and robust multicollinearitydianostic methods using Monte Carlo simulations

2 Methodology

A Overall F-Test

The test for significance of regression in the case of multiple linear regressions is carried out using the analysis of variance. The test is used to check if a linear statistical relationship exists between the response variable and at least one of the predictor variables. The statements of the hypotheses are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_j = 0$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j \quad (1)$$

The test for H_0 is carried out using the following statistic

$$F = \frac{MSR}{MSE} \quad (2)$$

Where MSR is the regression mean square and MSE is the error mean square. If the null hypothesis is true, then the statistic F follows the F distribution with k degrees of freedom in the numerator and n-(k+1) degrees of freedom in the denominator. The null hypothesis is rejected if the calculated statistic F is such that

$$F > f_{\propto,k,n-(k+1)} \quad (3)$$

B Individual T-Test

The t test is used to check the significance of individual regression coefficients in the multiple linear regression models. The hypothesis statements to test the significance of a particular regression coefficient, β_j, are

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0 \quad (4)$$

The test statistic for this test is based on the t distribution

$$t = \frac{\beta_j}{s(\beta_j)} \quad (5)$$

We would fail to reject null hypothesis if the test statistic lies in the acceptance region

$$-t_{\frac{\propto}{2},n-2} < t < t_{\frac{\propto}{2},n-2} \quad (6)$$

The test is significant if we have enough evidence to reject null hypothesis.

If some or all of the individual t-statistic are not significant, but the entire model is significant, then the multicollinearity may be influencing the model (Habshah et.al[4]). If there are several variables in the model, though, and not all are highly correlated with the other variables, this alone may not be enough. We could get a mix of significant and insignificant results disguising the fact that some coefficients are insignificant because of multicollinearity.

C Robust Variance Inflation Factor (RVIF)

In this study, we develop RVIF which is based on robust coefficient determination namely the $RR^2(MM)$. The $RR^2(MM)$ can be obtained from robust library of SPLUS software. It can be calculated as follows:

If the corresponding coefficient estimates computed using the final M-estimates, it can be obtained through final M estimator $\hat{\beta}$. If an intercept term $\hat{\beta}$ is included in the model, then the $RR^2(MM)$ is defined as

$$RR^2(MM) = \frac{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}}) - \sum \rho(\frac{y_i - \hat{y}_i}{\hat{s}})}{\sum \rho(\frac{y_i - \hat{\mu}}{\hat{s}})} \quad (7)$$

Where $\hat{\mu}$ is the location M estimate corresponding to the local minimum of

$$Q_y(\mu) = \sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) \quad (8)$$

Such that

$$Q_y(\mu) < Q_y(\mu^*) \quad (9)$$

Where $\hat{\mu}$ is the sample median estimate. can also be obtained from coefficient determination of MM estimators algorithm. The RVIF which is defined by replacing R^2 in VIF with $RR^2(MM)$ is called the RVIF(MM).

D OLS Estimates

The least square method is a very popular technique which is used to compute estimations of parameter and to fit data. It is one of the oldest techniques modern of statistics as it was first published in1805 by French mathematician Legendre in a now classic memoir. But this method is even older because it turned out that, after the publication of Legendre's memoir, Gauss another famous German mathematician, published another memoir (1809) in which he mentioned that he had previously discovered this method and used it as early as in1795. Nowadays, the least square is widely used to find and estimate the numerical values of parameters to fit a function to a set of data and to characterize the statistical properties of estimates.

The oldest and most frequent use of OLS consists of adjusting a model function to best fit a data set. The OLS equation is written in a form similar to the simple regression case:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (10)$$

Where

Y_i is the value of the response variable in the *i*th trial
$\beta_0, \beta_1, \beta_2$ are the parameters
X_{i1}, X_{i2} are the known constant, namely, the value of the predictor variables in the *i*th trial
ε_i is a random error term with mean 0 and variance σ^2

To express in matrix term, the general linear model is

$$Y = \beta X + \varepsilon \quad (11)$$

Where

Y is a vector of responses
X is a matrix of constants
β is a vector of parameters
ε is a vector of independent normal variable with mean 0 and constant variance

The OLS estimator for β is $b = (X'X)^{-1}X'Y$ and the sum of squares of residuals and total sum of square are $Y'Y - b'X'Y$ and $Y'Y - \left(\frac{1}{n}\right)Y'JY$ respectively.
Just as in the simple regression case, the R-squared is defined to be

$$R^2 = 1 - \frac{SSE}{SST} \tag{12}$$

And it is interpreted as the proportion of the sample variation in Y_i that is explained by the OLS regression line. By definition, R^2 is a number between 0 and 1.

E M-estimates

M estimators were first proposed by Huber [5] and are based on minimising a function of the residuals

$$\min \sum \rho(r_i) \tag{13}$$

Where ρ is a symmetric function with a unique minimum at zero. By differentiating the equation with respect to parameters yield,

$$\sum \psi\left(\frac{r_i}{\hat{\sigma}}\right) x_i = 0 \tag{14}$$

Where ψ is the derivative of ρ and the residuals have been scaled. However, this estimator does not achieve high breakdown point.

F S estimates

Least median squares and least trimmed squares are defined by minimizing a robust measure of the scatters of the residuals. Yohai and Rousseeuw [9] generalized these as S estimators that minimize the dispersion of the residuals,

$$\hat{\beta}_s = \arg\min \hat{\sigma}(\beta) \tag{15}$$

Where the dispersion $\hat{\sigma}(\beta)$ is the solution of $\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right) = b$ where $\rho$is replaced by an appropriate weight function.

G MM-estimates

It has been proven that robust regression estimators are more reliable and efficient than least squares estimators especially when disturbances are non normal. “Non normal disturbances” are disturbance distributions that have heavy or fatter tails than the normal distribution and are prone to produce outliers. Since outliers greatly influence the estimated coefficients, standard errors and test statistics, the usual statistical procedure may be the most inefficient as the precision of the estimator has been affected.

A better approach is to consider the robust procedure. This procedure fit a regression by using estimators that dampen the impact of influential points and then to detect outliers which are points lying far away from the pattern formed by the good points and has large residuals from the robust fit.

MM estimation, introduced by Yohai[10], combines high breakdown estimation and M estimation. It has both the high breakdown property and higher statistical efficiency than S estimation. MM estimators are in fact M estimators with an auxiliary scale estimate obtained from S estimator. MM estimators combine high efficiency with high breakdown and require the same computation time as S estimator. The MM estimators are defined in three stages where the first and the second stage are to achieve high breakdown point and the third stage is to aim for a high efficiency.

I. In stage one, to compute an initial regression estimate β_0 which is consistent robust with high breakdown point, possibly 50% using the S estimator.
II. In second stage, compute the residuals of the initial estimate,

$$r_i(\beta_0) = y_i - \beta_0 X_i', \quad 1 \leq i \leq n \tag{16}$$

Then compute an M estimate of errors scale $\hat{\sigma} = \sigma(r(\beta_0))$. $\hat{\sigma}$ is a solution of $\frac{1}{n}\sum \rho_0 \left(\frac{r_i}{\hat{\sigma}}\right) = b$ such that $\frac{b}{a} = 0.5$ where a is maximum ρ_0.

(Using a function ρ_0 satisfying Huber M estimation assumptions)

III. The third stage is to compute an M estimate of regression parameters based on a proper redescending psi-function. To find a solution $\hat{\beta}$ using an iterative procedure starting $\widehat{\beta_0}$.

Let $\rho_1$be another function satisfying Huber M estimation assumptions such that $\rho_i \leq \rho_0$, the influence function denoted as $\psi(t)$ is obtained by differentiating the objective function $\rho(t)$, that is $\psi(t) = \rho'(t)$. There are several functions of $\rho(t)$ and $\psi(t)$ to choose from, and in this study, Huber influence function will be employed.

Then, the MM estimator β_1 is defined as any solution of $\sum \psi_1 \left(\frac{r_i}{\hat{\sigma}}\right) x_i = 0$ and this equation must satisfy $Q(\beta_1) \leq Q(\beta_0)$ where $Q(\beta) = \sum \rho_0 \left(\frac{r_i}{\hat{\sigma}}\right)$ and $\rho_1(0/0)$ is defined as 0.

S Plus contains the lmRobMM function, which generates the above procedures. The MM estimation method in S Plus has a variety of statistics for diagnostics and inference including p-values, R-squared values, test for bias and residual scale estimates.

3 Analysis

In order to investigate the effect of high leverage points on multicollinearity pattern of the data, data set which was introduced by Kutner et al.[7] is considered. Commercial Properties data containing 81

observations is taken from the suburban commercial properties. The response variable is rental rates which were regressed to the age (X_1), operating expenses and taxes (X_2), and vacancy rates (X_3).

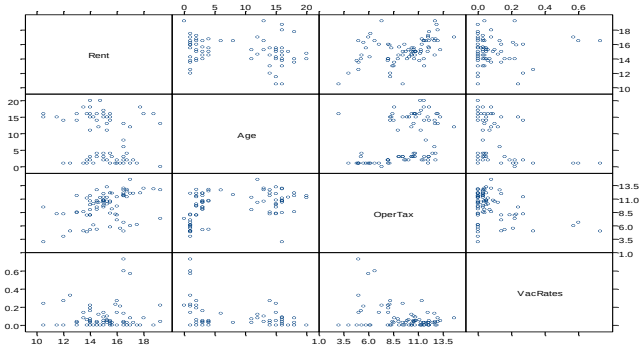

Figure 1 :Scatter Plot of Original Commercial Properties Data Set

According to Figure 1, none of the explanatory variables in the original data set is collinear since there is no linear pattern exists between any pair of explanatory variables. Also, we can see that there is no outlying point in the scatter plot. Hence, no outliers and automatically no high leverage point appears in this data set.

For the original data, set the F-test is significant with F-value=17.53 which is greater than F(0.95;3,77) = 2.74 and p-value=0.00 which indicates that there exists linear relationship between the variables in the model. The rental rates are related to age, operating expenses and taxes and vacancy rates. The value of R-squared is 0.4058 and 0.3833 for OLS method and Robust method respectively. Meanwhile, the value of residual standard error σ , for OLS is 1.351 and the value of residual scale estimate, σ , for Robust method is 0.9261. When we compare the value σ for both methods, we can obviously see that Robust method provides smallervalue σ . Hence, the Robust method is better than OLS method in order to make prediction.

Table 1 :VIF's for original Commercial Properties Data Set

	Original data set	
Variables	CVIF	RVIF
Age	1.196272	1.205786
OperTax	1.308633	1.314210
VacRates	1.186533	1.064254

The results in Table 1 indicate that when this data set doesn't contain any collinearity enhancing observations, both the classical VIF and RVIF(MM) are not exceeding their cutoff points. Thus, these results confirm that this data set is non collinear data set. No multicollinearity problem presents in the data set.

This data set has been modified to have high leverage observations. In order to modify this data set, the first observation of the first two explanatory variables is replaced with a large value of high leverage point which is 300.

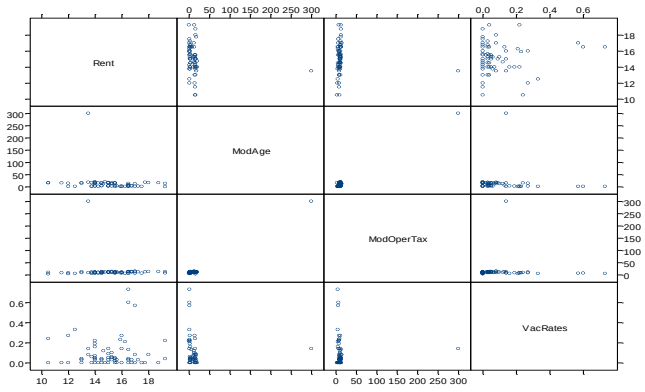

Figure 2 :Scatter Plot of Modified Commercial Properties Data Set

As soon as the data is modified, the added high leverage points in X_1 and X_2 pull the regression line towards themselves and change the collinearity pattern of the data which obviously can be seen in Figure 2. This indicates that a large value of high leverage point has changed the collinearity pattern of the data. It is also shown that there exists an outlier which is high leverage point in the data set.

After modifying the data set, the F-test is also significant with F-value= 6.85 which is greater compared to F(0.95;3,77) = 2.74 and p-value=0.00 β_3(VacRates) for the OLS is not significant with t-value =0.1778 which is less than t(0.975, 77)= 1.99 and p-value = 0.8594. This behavior is an indicator for the existence of multicollinearity in the data set (Montgomery et al.,[8]; Kutner et al.,[7]; Chatterjee and Hadi,[3]). The consequences of multicollinearity make OLS estimates and their standard error sensitive to small changes in the data. The data results will not be robust. However, the result also suggested that the robust method coefficient determinations are significant. These indicate that the robust method fits the model and resistant to the high leverage points and the parameter estimates will not be affected by these points.

Table 2: VIFs for modified Commercial Properties Data Set

	Modified data set	
Variables	CVIF	RVIF
Age	29.77288	1.198160
OperTax	29.78436	1.302256
VacRates	1.01283	1.064252

As shown in Table 2, the classical VIF indicates severe multicollinearityafter the data are modified by creating high leverage points which cause collinearity in the data set. However, our proposed RVIF(MM) is resistant to this added high leverage points and doesn't show collinearity for the data compared tothe CVIF which indicate severe multicollinearity. It is evident from the results that the high leverage points are the source of multicollinearity in the data set. The results reveal that high leverage points which are claimed by Kamruzzaman and Imon [6] to cause multicollinearity in non-collinear data set, is based on the classical VIF which is not resistant to these points.

A simulation study is conducted to further asses the performance of our new proposed Robust VIF. Three explanatory variables were considered in which each variable was generated from N(0,1). We refer to this generated data as the clean independent variables. In order to create collinearity enhancing observations, the clean data is replaced by certain percentage of high leverage points. The level of high leverage points varied from 0% to 20%. The sample size will also be varied which are

20, 40, 60, and 100 with 1000 replications in each simulation run. In order to obtain collinearity enhancing observations, the high leverage points were replaced in all three explanatory variables. We contaminate the independent variables with N(20,9). The average values of classical and Robust VIF were computed over 1000 simulation runs.

The result shows that in normal situations, the values of the RVIF(MM) are close to the classical VIF which indicate that they are as good as CVIF in diagnosing the collinearity pattern of the data correctly. As to be expected, when there is not any high leverage points in the data set, the CVIF confirmed that the data set is not collinear. It can be observed that by increasing the percentage of high leverage points, the value of CVIF become larger. Hence, the CVIF is very sensitive to the presence of added high leverage points to the data set. However, the value of RVIF(MM) doesn't change drastically and consistently close as we increase the percentage of high leverage points. It is evident from the results that RVIF(MM) is not affected by the increased in the percentage of high leverage points. The results also point out that the RVIF(MM) always robust against high leverage collinearity enhancing observations. The results of simulation study agree reasonably well with the results of real data.

4 Conclusion

In this study, we proposed a robust RVIF(MM) for detecting the source of multicollinearity which is caused by high leverage points in the data set. Recently, high leverage points are known to be another source of multicollinearity. The results of the study signify that the high leverage points has an unduly effect on the classical multicollinearity diagnostics, specifically the VIF. In this situation, sometimes the classical VIF conveys the misleading interpretation about collinearity pattern of the data. The results of the real data and simulation study reveal that high leverage points are the source of multicollinearity evidently by the failure of the RVIF in diagnosing multicollinearity; these contradict the results of CVIF which indicate existence of multicollinearity in the data set. The simulation studies clearly show that the robust MM estimator offers the most feasible option over other estimators when multicollinearity present. Another important conclusion is that the RVIF(MM) is the most resistant diagnostic measure to high leverage points. Robust regression estimation technique should perform as well as OLS when no multicollinearity problem exists, andit performs much better than OLS when multicollinearity problem exists because of high leverage point.

References

[1] Belsley.D.,Kuh.E.,&Welsh.R.(1980). Regression Diagnostics: Identifying Influential data and Sources of Collinearity. Wiley, New York.
[2] Chatterjee.S.,Hadi.A.(1988). Sensitivity Analysis in Linear Regression. Wiley, New York.
[3] Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn., Wiley, New York.
[4] Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Statist.
[5] Huber, P.J. (1964). Robust Estimation of a Location Parameter, Ann. Math. Statist.
[6] Kamruzzaman.M.,&Imon.A.H.M.R.(2002). High Leverage Point: Another Source of Multicollinearity. Pak. J. Statist.
[7] Kutner.M.H.,Nachtsheim.C.J, &Neter.J(2004). Applied Linear Regression Models. 4thEdn. McGraw Hill.
[8] Montgomery.D.,Peck.E.,&Vining.G.G(2001). Introduction to Linear Regression Analysis. 3rdEdn., Jon Wiley and Sons, New York.
[9] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. Robust and Nonlinear Time Series Analysis, Lecture Notes in Statist.
[10] Yohai.V.J.(1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. Ann. Statist.